Article:
Handley, Zoe Louise orcid.org/0000-0002-4732-3443 (Accepted: 2014) Constructing an evidence-base for future CALL design with 'engineering power' : The need for more basic research and instrumental replication. EUROCALL Review. ISSN 1695-2618 (In Press)

Article:

Constructing an evidence-base for future CALL design with 'engineering power': The need for more basic research and instrumental replication

Zöe Handley
Department of Education, University of York, UK

zoe.handley @ york.ac.uk

Abstract

This paper argues that the goal of Computer-Assisted Language Learning (CALL) research should be to construct a reliable evidence-base with 'engineering power' and generality upon which the design of future CALL software and activities can be based. In order to establish such an evidence base for future CALL design, it suggests that CALL research needs to move away from CALL versus non-CALL comparisons, and focus on investigating the differential impact of individual attributes and affordances, that is, specific features of a technology which might have an impact on learning. Further, in order to help researchers find possible explanations for the success or failure of CALL interventions and make appropriate adjustments to their design, it argues that these studies should be conducted within the framework of Second Language Acquisition (SLA) theory and research. Despite this, a recent review of research examining the effectiveness of CALL in primary and secondary English as a Foreign Language (EFL) found that CALL vs. non-CALL comparisons are still common and studies focusing on individual coding elements are rare. Further, few studies make links with SLA and few measure linguistic outcomes using measures developed in the field of SLA. One reason for this may be poor reporting of methods and difficulty in obtaining the instruments used in SLA research. Reporting guidelines and the use of the IRIS database (www.iris-databse.org) are introduced as possible solutions to these problems.

Keywords: Research methods, basic research, second language acquisition, replication, instruments.

1. Introduction

More basic Computer-Assisted Language Learning (CALL) research —and replications thereof— is required to permit researchers to construct a reliable evidence-base with 'engineering power' for the design of future CALL software and activities. An evidence-base with 'engineering power' is one that is sufficiently specific that it translates into CALL designs which work in practice (Burkhardt and Schoenfeld, 2003). Basic research refers to studies which provide insights into what specific features of digital environments create conditions and engage learners in processes that promote Second Language Acquisition (SLA), as well as what task variables promote SLA (Pederson, 1987).

My experience of synthesizing the literature in the field (Macaro, Walter & Handley, 2012), however, suggests that CALL research like educational research (Burkhardt & Schoenfeld, 2003) and SLA research (Porte, 2013) more broadly, is failing to achieve this and what we have instead is an accumulation of studies whose findings cannot easily be connected to those of other studies in the broader field of SLA or even within CALL itself.

Firstly, broad atheoretical CALL versus non-CALL comparisons of 'pen-and-paper' versus 'traditional' classroom activities are still common in the CALL evidence-base (Macaro et al., 2012) despite Pederson's (1987) call for them to "forever be abandoned" (p. 125). There are two reasons that this is a concern. First such studies do not have 'engineering power' because they fall into the trap of equating medium with method (ibid.). That is, they fail to acknowledge the fact that a particular technology might be used in a variety of different ways to support language learning and to implement a variety of different approaches to and methods of language teaching (Garrett, 1991) and fully exploit the added value of new technologies (Yildiz & Atkins, 1993): "technology is often used to change and expand the intended learning outcomes rather than to increase the level of performance in exactly the same areas as those targets by classroom instruction" (Chapelle, 2010, p. 70). With respect to the latter, in CALL research the possibility to engage in language learning activities 'anytime, anywhere' through the use of mobile technologies has been exploited to implement spaced vocabulary learning (Lu, 2008) and to contextualise vocabulary learning, that is, adapt it to the learners' immediate local environment (Chen & Li, 2010; Hwang & Chen, 2013; Gutiérrez-Colon et al., 2013).

Broad atheoretical CALL versus non-CALL comparisons also do not have explanatory power: the experimental condition often differs in multiple ways from the control condition and it is consequently not possible to determine to which feature of the software any observed differences should be attributed. O'Hara & Pritchard's (2008) evaluation of the impact of preparing a hyperlinked multimedia PowerPoint report on students' breadth of vocabulary knowledge illustrates this point well. In this study, production of PowerPoint reports with access to on-line resources was compared with production of pen-and-paper reports with access to paper-based classroom resources. The experimental condition, in other words, differed in two ways from the control condition: the medium in which the report was produced (PowerPoint vs. pen-and-paper) and access to resources (online vs. classroom). It is impossible therefore to know whether the higher levels of vocabulary knowledge observed in the experimental group should be attributed to the medium in which the report was produced or to access to online resources.

Secondly, the majority of CALL research is not grounded in SLA theory (Macaro et al., 2012). Grounding CALL research in SLA theory helps researchers to identify possible explanations for the effectiveness of particular manipulations of CALL environments and subsequently make appropriate adjustments to their design to better support language acquisition (Pederson, 1987).

Thirdly, the outcome measures employed in many CALL studies were developed for the specific purposes of the study in question and often differ from those commonly used in SLA research (Macaro et al., 2012). A combination of multiple-choice questions and ratings of learners' certainty in their choices was used as a measure of fluency of lexical recall in a study investigating the effects of different combinations of multimedia presentation on vocabulary learning (Kim & Gilman, 2008) rather than more widely accepted measures such as response latency, i.e. reaction time, for example. This is problematic because failure to engage in instrumental replication, i.e. to use the same outcome measures as employed in previous research, limits the comparability of studies

(Polio, 2012) and coherence of the discipline (Burkhardt & Schoenfeld, 2003), and is a barrier to meta-analysis (Oswald & Plonsky, 2010; Slavin, 1995). Aggregating the results of quantitative studies, meta-analyses are a key tool in the construction of an evidence-base within a discipline. When outcome measures are not consistently operationalised, they, however, produce less reliable estimates of effects and are challenging to interpret (ibid.). It has therefore been suggested that the principle inclusion criterion for a meta-analyses ought to be the construct validity of measures of the dependent variable: "a meta-analysis focusing on school achievement as a dependent measure must explicitly describe what is meant by school achievement and must only include studies that measure what is commonly understood as school achievement" (Slavin, 1995, p. 13), for example.

Finally, methods are frequently not adequately reported to permit replication (Macaro et al., 2012). In particular, instruments are often not provided (ibid.). Replication is, however, a cornerstone of scientific enquiry, necessary to ensure the construction of a reliable evidence-base (Polio, 2012) which has generality (Burkhardt & Schoenfeld, 2003). Reliability refers to the extent to which the individual findings have been validated through follow-up studies. Generality (or generalizability) refers to the extent to which individual findings have been demonstrated to hold in a wide range of contexts (ibid.). The demonstration of generality is perhaps the most important motivation for replication in CALL and SLA more broadly given the range of contextual variables that might have an impact on language learning (Chun, 2012).

In summary, current approaches to CALL research "are encouraging an *accumulation* of vaguely inter-connected research findings rather than the *construction* of knowledge across independent studies" (Porte, 2013, p. 12, original emphasis) which can be translated into designs for future CALL software and activities. In response to this, in the remainder of this paper, I introduce some of the different forms that basic research and replication might take within the field of CALL, and introduce IRIS (www.iris-databse.org), a digital repository of instruments, materials and stimuli used to elicit data in peer-reviewed research into second and foreign languages, as a resource to facilitate replication and promote the design of comparable studies. First, however, it is necessary to introduce the concept of 'engineering power'. Where possible, as above, all ideas will be illustrated with examples drawn from Macaro et al.'s (2012) systematic review of research on the use of technology in primary and secondary English as a Foreign Language (EFL) teaching.

2. 'Engineering power'

Like the automotive engineer designing and tuning a Formula 1 racing car, the early CALL researcher designing and optimising a learning environment was faced with a myriad of design options: "how and when to use graphics, sound feedback, branching from one learning task to the next based on learner response or request for new material, and how to display all these coding options accurately and efficiently" (Pederson, 1987, p. 100). To that list today we can add: how and when to provide interaction with other learners and the teacher (e.g. synchronously or asynchronously, one-to-one or many-to-one), how and when to personalise learning (e.g. based on attainment or context/location), and so on. The problem is that the theories that CALL researchers have to draw on such as socio-cultural theory are not sufficiently constrained —do not specify under what conditions the theory applies— and specific to translate into designs for CALL software and activities that work in practice (Burkhardt & Schoenfeld, 2003). In the same way that medium does not equate to method and there are many different ways in which a single technology might be employed to facilitate language learning (see above), there are often many different ways in which a particular theory might be translated into designs for CALL software and CALL activities.

In most studies within the field of CALL, socio-cultural theory has been translated into designs which exploit technology to provide learners access to more able partners (see for example Lund, 2008 and Sasaki & Takeuchi, 2010). It has, however, also been argued that support might be provided through access to appropriate resources as well as access to more able partners (Luckin & Clark, 2011; van Lier, 2004), for example. Grand theories such as socio-cultural theory are therefore not adequate to guide the design of CALL software and activities (Burkhardt & Schoenfeld, 2003). Highly specified 'local' theories which take into account the skill (reading, writing, speaking, or listening) or knowledge (vocabulary, grammar or pronunciation), the learner and the learning context, are rather what is required. In other words, like the 'craft' knowledge that practising teachers construct, such theories would be concrete, contextually rich and linked with practice (Hiebert, Gallimore & Stigler, 2002).

Further, to have "engineering power", the CALL evidence-base needs to have *generality*, that is, "go beyond the specific environment being examined, in order to make a contribution to knowledge of affordances of a technology or language learning processes" (Stockwell, 2012, p. 154). This will only be achieved if we abandon broad CALL versus non-CALL comparisons and focus our research efforts on attributes and affordances which transcend multiple specific technologies (Colpaert, 2010; Pederson, 1987). Attributes refer to features of the computer which have the potential to support and develop cognitive processing, such as symbol systems, multimedia and random access (Colpaert, 2010; Pederson, 1987). Affordances are features of the computer which enable learners to engage in processes that support language learning (Colpaert, 2010). These include the possibility to access authentic materials and interact with individuals and groups in the target language (ibid.). Kim and Gilman's (2008) systematic examination of the differential impact of different combinations of multimedia on learners' retention of vocabulary is a good example of a study with engineering power. It is specific and examines the impact of attributes which transcend a wide variety of technologies.

3. Basic CALL research

More basic CALL research is, however, required to allow us to *construct* an evidence-base upon which to design future CALL. Basic research refers to studies designed "to discover something about how students best learn a language", i.e. which "provid[es] explanatory data and add[s] to the theoretical bases for second language learning" (Pederson, 1987, p. 125). In other words, basic CALL research goes beyond evaluation and asks "Why did it work?" in addition to "Did it work?" (Levy & Stockwell, 2006, p. 42) and draws on and contributes to the development of SLA theory. Engaging in basic research, it has been suggested, has two benefits. First trials of complex health education interventions suggest that interventions grounded in appropriate theory are more likely to be effective (Campbell, Fitzpatrick, Haines et al. 2000). Second, where trials are unsuccessful, theory helps researchers identify possible explanations for failure to achieve learning goals and refine the design of interventions, in this case CALL software and activities (Pederson, 1987).

Basic CALL research has tended to take one of three forms: (1) exploratory research, (2) observational research, or (3) narrowly focused experimental research. Exploratory research is characterised by ethnographic studies in which researchers observe and interview students about their naturalistic use of CALL software with a view to generating theories regarding what features of digital environments create conditions and engage learners in processes that promote SLA (Pederson, 1987). An example of an informative ethnographic study is Gruber-Miller & Benton's (2001) examination of the *VRoma MOO* (1) for Latin. In observational studies the processes that students engage in during software use are logged and the relationship between software use and

learning gains is explored. An observational study with engineering power would resemble Proctor, Dalton and Grisham's (2007) investigation of native speakers' and English Language Learners' use of the *Universal Literacy Environment*, but track learners use of the different scaffolds provided at a more fine-grained level than overall frequency of use of scaffolds. Narrowly focused experimental studies isolate out the specific attributes and affordances of a technology which might have a differential impact on learning, and explore hypotheses grounded in SLA theory and research. In other words, narrowly focused experimental studies explore "the relative effectiveness of the pedagogical techniques that [a particular technology] implements, i.e., different types of feedback, online help, textual annotations, glossing formats, etc." (Burston, 2006, p. 258). Kim & Gilman's (2008) investigation of the differential impact of different combinations of multimedia on vocabulary knowledge is a good example of a narrowly focused experimental study. Another example is Dalton et al.'s (2011) comparison of different versions of a reading tutor integrating different forms of support, namely vocabulary versus reading support. It should, however, be noted that both vocabulary support and reading support could be realised in a number of different ways.

All of the above methods have the potential to make a significant contribution to our understanding of the conditions and processes which support SLA, as long as researchers engage with SLA theory and instrumental replication (see below). They are, however, not without their critiques. The value of narrowly focused experimental studies in particular has been questioned:

The treatment method leads to a danger that all experiments with computers and learning will be failures: either they are trivial because very little happened or they are "unscientific" because something real did happen and too many factors changed at once. (Papert, 1987, p. 26)

Two 'egineering' approaches to educational research are therefore beginning to attract attention in the field of CALL. These are design-based research (Barab & Squire, 2004; Burkhardt & Schoenfeld, 2003; Yutdhana, 2008) and educational engineering (Colpaert, 2006, 2010). In contrast, with the scientific approach to research upon which conventional methods draw, the engineering approach is transformative. That is, engineering research, like much educational research, is practice-oriented and aims to both understand "how the world works" and "help it to work better" (Burkhardt & Schoenfled, 2003, p. 5). It achieves this by "us[ing] existing knowledge in experimental development to produce new or substantially improved materials, devices, products, and processes including design and construction" (Higher Education Research Funding Council, 1999, p. 4). Design-based research (also referred to as design experiments and design research) refers to an approach in which 'local' theories of learning and teaching are tested and refined through iterative cycles of design and evaluation in collaboration with end-users, i.e. learners and teachers, and gradually scaled up and rolled out for use in practice (Barab & Squire, 2004; Burkhardt & Schoenfeld, 2003; Gorard, Roberts & Taylor, 2004; Yutdhana, 2005). In other words, in addition to being transformative and practice-oriented, design-based research recognizes and values teacher cognition (see Borg, 2003; Kumaravadivelu, 1994) and is impact-oriented. A study adopting Bannan-Ritland's (2003) Integrative Learning Design (ILD) framework for design-based research, for example, would begin with informed exploration of the learning context and problem, i.e. needs analysis. This phase of the design process, which would also include a review of the literature and identification of appropriate learning theory, would result in a specification for the design of the CALL system. In the next phase of the design process, enactment, would involve translating the requirements to a design and developing a prototype. The local impact of the design would then be evaluated in the next phase, the results of which would lead to adjustments to the design and further

cycles of evaluation. Design-based research aims to produce 'shareable theories' (Design-Based Collective, 2003, p. 5). In the final stage, others would therefore be encouraged to adopt the design and theory to allow evaluation of broader impact. Pardo-Ballester & Rodriguez's (2009, 2010) development of online readings for elementary learners of Spanish for business and engineering, for example, is grounded in design-based research.

Educational engineering as conceived by Colpaert (2006, 2010) is also characterised by iterative cycles of development. The approach, however, is grounded in theories of motivation and the assumption that CALL ought to "support the learner in better achieving learning goals" (Colpaert, 2010, p. 273) and prioritises process over product as outcome measures: "Engineering does not focus on measurable significant differences on a product level, but rather on observable phenomena on a process level" (ibid., p. 262). The departure for design and research within this approach is therefore an examination of learner goals. Having identified learner goals through focus group discussions and compared them with other competing goals and in particular pedagogical goals, appropriate learning theories to operationalise the competing goals are identified and a design for the CALL software and tasks is elaborated. The resulting design and any design and theoretical questions that arise from it are then explored through iterative cycles of design and evaluation as in design-based research. Educational engineering has therefore been characterised as 'slow research' and all of the projects that have adopted this research to date are still on-going. It is therefore not possible to discuss any completed projects at this point. For a list of on-going projects see Colpaert (2010).

In summary, whatever methodology is adopted, drawing links with SLA theory and research is essential to drive the construction of an evidence-base for the design of future CALL software and activities forward. It will "lead to a stronger focus on the learning process rather than the technology" (Stockwell, 2012, p. 160) and, by providing insights into the reasons for the success and failure of CALL software and activities, it helps researchers and developers refine the design of future CALL software.

4. Replication in CALL

Replication is also required to *construct* a *reliable* evidence-base with *generality*. Exact replications, in which researchers attempt to copy the original study as closely as possible using identical subjects, conditions, and instruments, among other things, should be conducted where possible to allow the validation of findings (Polio, 2012; Porte & Richards, 2012). Instrumental replications, approximate replications in which the same outcome measures as used in previous research are employed, should be conducted in a range of different contexts to permit the demonstration of the generality of findings and also to permit comparisons and meta-analyses of studies within CALL and in the broader field of SLA (Polio, 2012). Further, conceptual replications in which findings are tested using a different study design, in particular different data collection procedures (e.g. observation versus self-report) are essential to demonstrate the validity of findings, i.e. to demonstrate that they are not artefacts of the original design (Polio, 2012; Porte & Richards, 2012).

Replication in CALL research, as in SLA research more broadly, has, however, largely been neglected, with the exception of a number of studies which have replicated findings of SLA research (Chun, 2012). In fact, some question whether replication is even possible in CALL given the pace of technological advances and the fact that older technologies quickly fall into obsolescence (Chun, 2012). This argument does not, however, hold if we move away from broad CALL versus non-CALL comparisons and

focus our research efforts on the exploration of the impact of attributes and affordances which transcend individual technologies, new and old, as discussed above.

A greater problem, however, is that, as in SLA research more broadly (Polio & Gass, 1997), CALL research is not adequately reported to permit instrumental replication, let alone exact replication (Macaro et al., 2012). Instruments, including background questionnaires, measures of proficiency, instruments for data elicitation and pre- and post-tests, and coding frameworks (Polio & Gass, 1997), are rarely provided in CALL studies, and often barely discussed in the methods sections of research articles (Macaro et al., 2012). While it is always possible to contact authors to request materials, researchers can be difficult to track –they move– and they may not always be able to easily locate materials within their archives (Marsden & King, 2013, Marsden & Mackey 2014).

One way to overcome these problems is to introduce reporting guidelines, as suggested by Polio & Gass (1997, p506):

[E]xamples of what might ultimately be useful to researchers [include]: (a) Detailed guidelines and examples of coding categories, (b) A listing of examples that were excluded from consideration, (c) Measures of proficiency (descriptions of tests where security is a problem), (d) Instruments for data elicitation, including pre-tests and post-tests, (e) Experimental protocols and instructions to subjects, and (f) Demographic background of subjects.

Experience in the health sciences has demonstrated that, in addition to permitting replication, the introduction of reporting guidelines has increased the quality of published research (Moher, Jones & Lepage, 2001). Researchers interested in building on Polio & Gass's (1997) suggestions are encouraged to consult the reporting guidelines for relevant forms of research in the health sciences, including CONSORT for randomized controlled trials (Moher, Schulz, & Altman, 2001; www.consort-statement.org), i.e. experimental research, STROBE for observational studies (www.strobe-statement.org), and COREQ for qualitative interview-based research (Tong, Sainsbury, & Craig, 2007).

Whether or not reporting guidelines are introduced, barriers to replication will, however, remain. First it will remain difficult to replicate studies which have already been published. Second, it will remain difficult to locate instruments. Articles in electronic databases are typically indexed, that is, assigned thesaurus terms, on the basis of the title and abstract alone. Even if we were to adopt reporting guidelines, it would simply not be possible to provide sufficient information to index instruments in the abstracts of CALL and SLA research articles (see below for a list of the dimensions on which it would be desirable to index instruments for use in CALL and SLA research).

5. The IRIS database

Instruments for Research into Second Language Learning and Teaching (IRIS) is an open access digital repository of materials used to collect data in research on second and foreign language acquisition developed and curated by the Digital Library at the University of York which might help researchers in the fields of CALL and SLA overcome those barriers. All instruments held on the database have been used to collect data for a peer-reviewed publication, i.e. a peer-reviewed journal or conference proceedings, an edited book or a successful doctoral thesis. The database is searchable along a number of dimensions including instrument type, linguistic feature, and learner proficiency, and materials can be downloaded and re-used, with most held under a Creative Commons derivatives allowed non-commercial share-alike licence. In other words researchers "can remix, tweak, and build upon this work non-commercially, as long as [they] credit the

creators of the instrument and license [their] new creations under identical terms" (www.iris-database.org).

It is also possible for researchers to upload their own instruments to the database for use by other researchers. In fact, 30 top ranking journal editors are now encouraging uploads, including the editors of the following SLA journals: *Applied Linguistics, Language Learning, Language Teaching, Studies in Second Language Acquisition* and *The Modern Language Journal,* as well as *Computer Assisted Language Learning Journal* and *System.* IRIS currently holds over 850 documents bundled into approximately 280 instruments. The coverage of the database is wide, with over fifty instrument types represented, including language background questionnaires, cloze tests, grammaticality judgement tests, and elicitation tasks, and over forty research areas, including motivation, processing instruction, and task-based interaction.

As a research area CALL is currently under-represented with only two instruments in comparison with morphosyntax (grammar) for which over 100 instruments have been uploaded. In line with current interests in computer-mediated task-based language learning, however, a variety of tasks are held on the database which might be re-used and adapted in this area of research. These include tasks designed to:

- Investigate learners' use of communication strategies -e.g. García Mayo's (2005) decision-making task.
- Elicit specific morphosyntactic forms -e.g. Mifka Profozic's (2012) picture description tasks for eliciting the French *passé composé* and *imparfait.*
- Examine the impact of task complexity on the extent to which focus on form or meaning -e.g. Révész's (2011) argumentative tasks.

Moreover, if your area of interest in CALL or SLA is not represented and there is an instrument that you would like to examine or re-use, it is possible to get the IRIS team (iris@iris-database.org) to track down the materials for you by placing a request through the IRIS database.

6. Conclusion

Current CALL research which is dominated by broad media comparisons has resulted in "an accumulation of vaguely inter-connected research findings" (Porte, 2013, p. 12). In order to *construct* a *reliable* evidence-base with *'engineering power'* upon which to base future CALL design, more basic research —and replications thereof— is necessary. Instrumental replication is particularly important to permit researchers to build on the findings of previous research. In order to permit such comparisons, CALL researchers in the field are encouraged to contribute instruments from their peer-reviewed publications to the IRIS database. With nearly 5000 downloads to date, 15000 hits on the site, and references to the publications in which the instruments have been used, having materials on IRIS increases the visibility of individual researcher's work. Integrating the option to request downloaders to leave their name and e-mail address, having materials on IRIS also permits researchers to track the impact of their research.

Acknowledgements

IRIS is developed and curated by the Digital Library at the University of York, and directed by Emma Marsden (York, UK) and Alison Mackey (Georgetown, USA / Lancaster, UK). It is funded by the Economic and Social Research Council and the British Academy. The author would like to thank the IRIS team for supporting attendance at EUROCALL 2014 where an earlier version of this paper was presented, and in particular Emma Marsden for her suggestions on how to improve the paper.

References

Bannan-Ritland, B. (2003). The role of design in research: The integrative learning design framework. *Educational Researcher, 32*(1), 21-24.

Barab, S. & Squire, K. (2004). Design-based research: Putting a stake in the ground. *Journal of the Learning Sciences,* 13(1), 1-14.

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe and do. *Language Teaching,* 36(2), 81-109.

Burkhardt, H. & Schoenfeld, A. H. (2003). Improving educational research: Toward a more useful, more influential, and better-funded enterprise. *Educational Researcher,* 32(3), 3-14.

Burston, J (2006). Working towards effective assessment of CALL. In Donaldson, P. R. & Haggstrom, M. A. (eds.). *Changing Language Education Through CALL*. London: Routledge, 249-270.

Campbell, M., Fitzpatrick, R., Haines, A., Kinmouth, A. L., Sandercock, P., Spiegelhalter, D. & Tyrer, P. (2000). Framework for design and evaluation of complex interventions to improve health. *British Medical Journal, 321*, 694-696.

Chapelle, C. (2010). The spread of computer-assisted language learning. *Language Teaching, 43*(1), 66-74.

Chen, C.-M. & Li, Y. L. (2010). Personalised context-aware ubiquitous learning system for supporting effective English as a second language. *TESOL Quarterly, 20*(1), 27-46.

Chun, D. (2012). Review article: Replication studies in CALL research. *CALICO Journal, 29*(4), 591-600.

Colpaert, J. (2006). Pedagogy-driven design for online language teaching and learning. *CALICO Journal, 23*(3), 477-497.

Colpaert, J. (2010). Elicitation of language learners' personal goals as design concepts. *Innovation in language learning and teaching,* 4(3) 259-274.

Dalton, B., Proctor, C. P., Uccelli, P., Mo, E. & Snow, C. E. (2011). Designing for diversity: The role of reading strategies and interactive vocabulary in a digital reading environment for fifth-grade monolingual English and bilingual students. *Journal of Literacy Research, 43*, 68-100

Design-Based Research Collective (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher,* 32(1), 5-8

García Mayo, M. (2005). Interactional strategies for interlanguage communication: Do they provide evidence for attention to form? In A. Housen & M. Pierrard (Eds.*), Investigations in instructed second language acquisition* (Studies on Language Acquisition Series). Mouton de Gruyter.

Garrett, N. (1991). Technology in the service of language learning: Trends and issues. *Modern Language Journal,* 75, 74-101

Gorard, S., Roberts, K. & Taylor, C. (2004). What kind of creature is a design experiment? *British Educational Research Journal,* 30(4), 577-590.

Gruber-Miller, J. & Benton, C. (2001). How do you say 'MOO' in Latin? Assessing student learning and motivation in beginning Latin. *CALICO Journal,* 18(2), 305-38.

Gutiérrez-Colon, M.; Gibert, M.I.; Triana, I.; Gimeno, A.; Appel, C. & Hopkins, J. (2003). Improving learners' reading skills through instant short messages: a sample

study using WhatsApp. *WorldCALL 2013 – CALL: Sustainability and Computer-Assisted Language Learning Conference Proceedings.* University of Ulster, pp.80-84.

Hiebert, J., Gallimore, R. & Stigler, J. (2002). A knowledge base for the teaching profession: What would it look like and how can we get one? *Educational Researcher,* 31(3), 3-15.

Higher Education Research Funding Council (1999). *Guidance on submissions research assessment exercise,* Paragraph 1.12. London: Higher Education Funding Council for England and Wales 1999.

Hwang, W. Y. & Chen, H. S. L. (2013). Users' familiar situational context facilitate the practice of EFL in elementary schools with mobile devices. *Computer-Assisted Language Learning,* 26(2), 101-125

Kim, D.; Gilman, D. A. (2008). Effects of Text, Audio, and Graphic Aids in Multimedia Instruction for Vocabulary Learning. *Educational Technology & Society*. 11 (3): 114-126.

Kumaravadivelu, B. (1994). The post-method condition: (E)merging strategies for Second/Foreign language teaching. *TESOL Quarterly,* 28(1), 27-48.

Levy, M. & Stockwell, G. (2006). *CALL Dimensions: Options and Issues in Computer-Assisted Language Learning.* London: Lawrence Erlbaum.

Lu, M. (2008). Effectiveness of vocabulary learning via mobile phones. *Journal of Computer Assisted Learning,* 24(6), 515-525.

Luckin, R. & Clark, W. (2011). More than a game: The participatory Design of contextualised technology-rich learning experiences with the ecology of resources. *Journal of e-Learning and Knowledge Society,* 7(3), 33-50.

Lund, A. (2008). Wikis: A collective approach to language production. *ReCALL*, 20(1), 35-54.

Macaro, E., Handley, Z. L., & Walter, C. (2012). A systematic review of CALL in English as a second language: Focus on primary and secondary education. *Language Teaching,* 45(1), 1-43

Marsden, E. & King, J. (2013). The Instruments for Research into Second Languages (IRIS) digital repository. *The Language Teacher,* 37(2), 35-38.

Marsden, E. J., & Mackey, A. (2014). IRIS: a new resource for second language research. *Linguistic Approaches to Bilingualism,* 4(1), 125-130.

Mifka Profozic, N. (2012). Oral corrective feedback, individual differences and L2 acquisition of French past tenses. (Unpublished doctoral dissertation). University of Auckland.

Moher, D., Jones, A. & Lepage, L. (2001). Use of the CONSORT Statement and quality of reports of randomized trials. A comparative before-and-after evaluation. *The Journal of the American Medical Association,* 285, 1992-5.

Moher, D., Schulz, K. ., Altman, D. (2001). The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomized trials. *The Journal of the American Medical Association,* 285, 1987-91.

O'Hara, & Pritchard, (2008). Hypermedia authoring as a vehicle for vocabulary development in middle school English as a second language classrooms. *Clearing House: A Journal of Educational Strategies, Issues, and Ideas,* 82(2), 60-65.

Oswald, F. L. & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics,* 30, 85-110.

Pardo-Ballester, C. & Rodriguez, J. C. (2009). Using design-based research to guide the development of online instructional materials. In Chapelle, C. A., Jun, H. G., & Katz, I. (Eds.). *Developing and evaluating language learning materials* (pp. 86-102). Ames, IA: Iowa State University.

Pardo-Ballester, C. & Rodriguez, J. C. (2010). Developing Spanish online readings using design-based research. *CALICO Journal,* 27(3), 540-553.

Pederson, K (1987). Research on CALL. In Smith, W. F. (ed.). *Modern media in foreign language education: Theory and implementation*. Lincolnwood, Illinois: National Textbook Company, pp. 99-131.

Polio, C. (2012). Replication in published applied linguistics research: A historical perspective. In Porte, G. (ed.). *Replication research in applied linguistics*. Cambridge: Cambridge University Press, pp. 47-91.

Polio, C. & Gass, S. (1997). Replication and reporting. *Studies in Second Language Acquisition,* 19, 499-508.

Porte, G. (2013). Who needs replication? *CALICO Journal, 30*(1), 10-15

Porte, G. & Richards, K. (2012). Focus article: Replication in second language writing research. *Journal of Second Language Writing, 21*(2012), 284-193.

Proctor, C. P., Dalton, B. & Grisham, D. L. (2007). Scaffolding English language learners and struggling readers in a universal literacy environment with embedded strategy instruction and vocabulary support. *Journal of Literacy Research,* 39(1), 71-93.

Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *The Modern Language Journal,* 95(4).

Sasaki, A. & Takeuchi, O. (2010). EFL students' vocabulary learning in NS-NNS e-mail interactions: Do they learn new words by imitation? *ReCALL*, 22(1): 70-82.

Slavin, R. E. (1995). Best evidence synthesis: An intelligent alternative to meta-analysis. *Journal of Clinical Epidemiology,* 48(1), 9-18.

Stockwell, G. (2012). Diversity in research and practice. In Stockwell, G. (ed.). *Computer-Assisted Language Learning: Diversity in Research and Practice*. Cambridge: Cambridge University Press, pp. 147-163.

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus-groups. *International Journal for Quality in Health Care,* 19(6), 349-357.

van Lier, L. (2004). *The Ecology and Semiotics of Language Learning. A Sociocultural Perspective.* Boston: Kluwer Academic

Yildiz, R. & Atkins, M. (1993). Evaluating multimedia applications. *Computers in Education,* 21(1/2), 133-139.

Yutdhana, S. (2008). Design-based research in CALL. In Egbert, J. L. & Petrie, G. M. (eds.). *CALL research perspectives*. Mahwah, NJ: Lawrence Erlbaum, pp. 169-178.

Notes

[1] MOO stands for Multi-user Object Oriented domains. A MOO is an online 2D or 3D virtual world.